PHÁT TRIỂN HỆ THỐNG DỊCH ĐA NGỮ GIỮA TIẾNG VIỆT VÀ MỘT SỐ NGÔN NGỮ KHÁC

TỪ NĂM 2020 ĐẾN NAY, TS. NGUYỄN VĂN VINH – CHỦ NHIỆM ĐỀ TÀI CÙNG NHÓM NGHIÊN CỨU VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN & ỨNG DỤNG CỦA KHOA CÔNG NGHỆ THÔNG TIN – TRƯỜNG ĐẠI HỌC CÔNG NGHỆ, ĐẠI HỌC QUỐC GIA HÀ NỘI VÀ MỘT SỐ ĐƠN VỊ KHÁC NHƯ TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG -ĐH BÁCH KHOA, TRƯỜNG ĐH KINH TẾ KỸ THUẬT HÀ NỘI, TRƯỜNG ĐH KHOA HỌC XÃ HỘI VÀ NHÂN VĂN... ĐÃ THÀNH CÔNG NGHIÊN CỨU ĐỀ TÀI CẤP QUỐC GIA KC-4.0.12/19-25 "PHÁT TRIỂN HỆ THỐNG DỊCH ĐA NGỮ GIỮA TIẾNG VIỆT VÀ MỘT SỐ NGÔN NGỮ KHÁC" VÀ ĐƯA HỆ THỐNG VÀO SỬ DỤNG TẠI WEBSITE ITRITHUC.VN.

THIÊN BÌNH

HỆ THỐNG DỊCH ĐA NGỮ GIỮA TIẾNG VIỆT VÀ MỘT SỐ NGÔN NGỮ HẠN CHẾ TÀI NGUYÊN

Hiện nay, do hội nhập quốc tế Việt Nam và nhiều nước trên thế giới có quan hệ ngoại giao, kinh tế. Do vậy, lượng văn bản song ngữ giữa Việt Nam và các nước khác ngày càng lớn. Trong khi đó, nguồn nhân lực phiên dịch viên khan hiếm, giá thành dịch thuật cao. Do đó, hệ thống dịch tự động từ tiếng Việt ra các ngôn ngữ khác và ngược lại rất cần thiết. Ngày nay, nhiều hệ dịch đa ngữ thương mại như Google, Microsoft, Facebook... nhưng các hệ dịch thương mại này chất lượng chỉ tốt cho các cặp ngôn ngữ chính và nhiều tài nguyên, lĩnh vực chung như Anh – Việt, Việt – Anh..., còn các cặp ngôn ngữ trên các lĩnh vực chuyên biệt và tài nguyên hạn chế như Việt – Lào, Việt – Khmer chưa đạt chất lượng như mong muốn. Ngoài ra, hầu hết việc dịch dựa trên nền tảng Cloud nên vấn đề bảo mật dữ liệu cũng

như tuy biến về hệ thống dịch gặp khó khăn, đặc biệt là các lĩnh vực an ninh, quốc phòng và các lĩnh vực chuyên ngành khác như y tế, giáo dục, pháp luật, ...

Xuất phát từ lý do này mà TS. Nguyễn Văn Vinh cùng nhóm nghiên cứu của một số đơn vị trong cả nước đã phát triển riêng hệ dịch đa ngữ tự động từ tiếng Việt. TS. Nguyễn Văn Vinh nhấn mạnh: "Điểm mạnh của hệ thống dịch này và phương pháp dịch hiệu quả cho các cặp tài nguyên hạn chế hoàn toàn như một framework dịch có thể sử dụng để dịch các cặp ngôn ngữ khác. Đặc biệt là tùy biến để hệ thống có thể dịch tiếng Việt với các tiếng dân tộc của Việt Nam với chỉ một lượng nhỏ dữ liệu song ngữ, điều này các hệ dịch thương mại và trực tuyến như Google, Bing hoàn toàn không hỗ trợ".

Chia sẻ về cách tiếp cận của đề tài, TS. Nguyễn Văn Vinh cùng nhóm nghiên cứu cũng gặp những khó khăn nhất định. Đặc biệt là nguồn ngữ liệu song ngữ là quan trọng nhất, vì vậy việc phát triển hệ thống phần mềm mở dịch văn bản đa ngữ, phát triển phương pháp thu thập kho ngữ liệu song ngữ cho các ngôn ngữ, xây dựng kho ngữ liệu song ngữ cho các cặp ngôn ngữ có tài nguyên hạn chế trở nên cần thiết. Đối với việc nghiên cứu về vấn đề này, TS. Nguyễn Văn Vinh đề cập đến thách thức đặt ra: "Các nghiên cứu chủ yếu áp dụng cho cặp ngôn ngữ Anh – Việt, cũng có một số nghiên cứu áp dụng xây dựng ngữ liệu Trung –Việt, Việt – Trung, Việt – Lào, Việt – Khmer. Tuy nhiên, các nghiên cứu này chủ

yếu xây dựng thủ công, chưa áp dụng việc xây dựng hệ thống tự động cho các kho ngữ liệu song ngữ áp dụng cho bài toán dịch". Vì vậy, TS. Nguyễn Văn Vinh và nhóm nghiên cứu nhận định, việc nghiên cứu phát triển các phương pháp tự động hỗ trợ trích rút những kho ngữ liệu song ngữ chất lượng cho các cặp ngôn ngữ có tài nguyên hạn chế là chủ đề cần thiết và có tính thực tiễn cao.

ỨNG DỤNG HỆ THỐNG TRÊN WEBSITE ITRITHUC.VN

Từ năm 2020 đến nay, nhóm nghiên cứu đã liên kết, phối hợp với các đơn vị nghiên cứu, hoạt động chuyên môn sâu về lĩnh vực xử lý ngôn ngữ Trung, Lào, Khmer và dịch máy. TS. Nguyễn Văn Vinh nhắc đến kết quả của đề tài với việc xây dựng hệ thống dịch máy đa ngữ với 4 cặp ngôn ngữ Việt – Anh, Trung, Lào, Khmer được triển khai trên cổng itrithuc.vn; xây dựng kho ngữ liệu song ngữ Việt – Trung chất lượng tốt, Việt – Lào và Việt Khmer; công cụ trợ giúp xây dựng cơ sở dữ liệu song ngữ và đơn ngữ; phát triển bộ công cụ dịch máy mã nguồn mở phục vụ cộng đồng nghiên cứu và phát triển... Sau khi triển khai hệ thống dịch thử nghiệm trên Itrithuc.vn, Trung tâm phát thanh tiếng Khmer của Đài tiếng nói Việt Nam (VOV) và một số đơn vị của Bộ Công An, Bộ Quốc Phòng nhóm nghiên cứu đã nhận được các phản hồi tích vực và hoàn toàn có thể đưa vào ứng dụng trong thực tế.

Khẳng định ưu thế về hệ thống dịch của đề tài so với các hệ thống dịch hiện nay (Google Translate, Bing Translate, ...), TS. Nguyễn Văn Vinh cho biết: "Chúng tôi làm chủ hoàn toàn công nghệ dịch, phát triển engine dịch từ đầu nên hoàn toàn có thể áp dụng cho các đơn vị có dữ liệu bảo mật và các cặp ngôn ngữ mà

các hệ dịch hiện nay không hỗ trợ như tiếng Việt-tiếng dân tộc (Tày, Mường, ...). Thêm vào đó chất lượng dịch các cặp ngôn ngữ như Việt-Lào và Việt-Khmer của hệ thống dịch này tốt hơn so với các hệ thống dịch có trên thị trường hiện nay".

Để có được kết quả nghiên cứu và thành công của sản phẩm, TS. Nguyễn Văn Vinh cùng nhóm nghiên cứu đã khắc phục nhiều khó khăn về thời gian, tiến độ và đặc biệt là dịch Covid-19 đã làm việc nghiên cứu gặp nhiều bất lợi. Tuy nhiên, để TS. Nguyễn Văn Vinh cho biết: "Việc thực hiện đề tài diễn ra đúng thời điểm dịch Covid-19 bùng nổ khắp thế giới và Việt Nam. Điều đó dẫn đến việc mua sắm thiết bị và triển khai thử nghiệm cũng gặp khó khăn. Tuy nhiên với sự nỗ lực của các thành viên trong các đề tài nhánh, sự phối hợp ăn ý giữa các nhóm cùng với việc tổ chức báo cáo tiến độ hàng tuần đã giúp nhóm nghiên cứu khắc phục được các khó khăn trong quá trình nghiên cứu".

Thời gian tới, nhóm nghiên cứu muốn tiếp tục triển khai và phát triển sản phẩm này để được sử dụng lâu dài và phổ biến trong xã hội. Vì vậy, nhóm nghiên cứu mong muốn nhận được sự hỗ trợ quảng bá từ Bộ Khoa học và công nghệ, ĐH Quốc Gia và Trường ĐHCN để sản phẩm dịch được xã hội và doanh nghiệp biết đến nhiều hơn. Bên cạnh đó, nhóm nghiên cứu cũng tiếp tục nghiên cứu để cải tiến thêm chất lượng của hệ dịch này dựa vào các công nghệ tiên tiến hiện nay như Mô hình ngôn ngữ lớn và tạo sinh (Generative AI) ChatGPT.